AN INFORMATION RETRIEVAL ALGORITHM TO EXTRACT INFLUENTIAL FACTORS

A project submitted to Dean of Research and Postgraduate Studies Office in partial
Fulfillment of the requirement for the degree
Master of Science (Information Technology)
Universiti Utara Malaysia

By
NABILAH FILZAH BINTI MOHD RADZUAN

PERMISSION TO USE

In presenting this project in partial fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this project in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence by the Dean of Postgraduate and Research. It is understood that any copying or publication or use of this project or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my project.

Requests for permission to copy or to make other use of materials in this project, in whole or in part, should be addressed to

Dean of Research and Postgraduate Studies
College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman
Malaysia

ABSTRAK (BAHASA MALAYSIA)

Terdapat banyak factor yang boleh digunakan untuk menilai prestasi syarikat dari rujukan lepas tetapi hanya satu jumlah faktor terhad diperlukan untuk dengan cekap menilai prestasinya. Tujuan kajian ialah untuk membangunkan satu algoritma yang boleh mengambil paling minimum menyediakan faktor-faktor yang boleh digunakan untuk menilai persembahan syarikat. Harga saham telah digunakan sebagai faktor bersandar. Faktor-faktor dikeluarkan diketahui sebagai factor berpengaruh kerana factor ini didapati mempunyai pengaruh kuat di harga saham. Objektif kajian adalah untuk mendapatkan satu faktor-faktor berpengaruh komprehensif membangunkan algoritma pengekstrakan yang boleh mengenal pasti faktor yang mempengaruhi , dan fackor sekarang yang mempengaruhi harga saham syarikat-syarikat. Data mengandungi faktor kewangan yang diperolehi dari dokumen kewangan syarikat bermasalah dan syarikat tidak bermasalah menyenaraikan di bursa saham. Algoritma pengekstrakan telah dibangunkan dan dilaksanakan menggunakan bahasa pengaturcaraan Matlab. Keputusan menunjukkan daripada 33 faktor, 5 faktor telah didapati set minimum menghendaki menilai persembahan syarikat. Ini ialah hutang, pelaburan, jumlah aset, pusing ganti aset , dan modal kerja. Algoritma telah diuji di set data lain keputusan untuk mengeluarkan lebih dari 70 peratus maklum balas positif. Ini menunjukkan yang algoritma mampu menghasilkan satu contoh yang baik. Algoritma pengekstrakan membangunkan menunjukkan yang faktor yang mempengaruhi influencial menghasilkan boleh digunakan kerana garis panduan bagi syarikat memantau dan mengatur strategi cara-cara untuk peningkatan perniagaan.

ABSTRACT (ENGLISH)

Past literatures showed that there are many factors that can be used to assess company's performance but only a limited number of factors are needed to efficiently assess its performance. The aim of the study is to develop an algorithm that can extract a minimum set of factors that can be used to assess companies' performances. Stock price was used as the dependent factor. The factors extracted are known as influential factors because these factors were found to have strong influence on the stock price. The objectives of the study were to obtain a comprehensive influential factors from past literatures, develop an extraction algorithm that can identify influencial factors, and present factors that influenced companies' stock prices. Data consisted of financial factors that were obtained from financial documents of distressed companies and non-distressed companies listed on a stock exchange. The extraction algorithm was developed and implemented using Matlab programming language. Results showed that out of 33 factors, 5 factors were found to be the minimum set needed to assess the companies' performances. These were debt, investment, total asset, asset turnover, and working capital. The algorithm were tested on other dataset and results produced more than 70 percent of positive feedback. This indicates that the algorithm was able to produce a good model. The extraction algorithm developed showed that influencial factors produced could be used as guideline for companies to monitor and strategize ways for business improvement.

ACKNOWLEDGEMENT

Firstly, praise to Allah S.W.T. for guiding and blessing with perseverance and strength to complete the project. Apart from the efforts of me, the success of the project depends largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project. The special thank goes to my helpful supervisor Dr. Faudziah Ahmad and Prof. Ku Ruhana Ku Mahamud. I can't say thank you enough for his tremendous support and help. Without her encouragement and guidance this project would not have materialized. My grateful thanks also go to my parents and siblings, who held faith in me and pushed me to succeed. A big contribution and supported from you is very great indeed. Special thanks also go to my friends those who supported and motivated me during the project completion was vital for the success of the project. Last but not least, I would like to thank to all University Utara Malaysia management especially College of Arts and Sciences staff and those who involved directly or indirectly in the project. May Allah bless all of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

KDD	KNOWLEDGE DISCOVERING IN DATABASE
EM	EXPECTATION MAXIMIZATION
CCR CURVE	CORRELATED COMPONENT REGRESSION CURVE

CHAPTER ONE

INTRODUCTION

1.1 Background

Influential factors are significant factors which can bring impact to the environment of studies. According to Ouwens et al. (2001), influencial is when a data structure has a large impact on the estimated model parameters and their characteristics are investigated to know how much they influenced the environment. The influential factors can be captured from large or small datasets or databases using some methods and represent useful information to a company. Gray (1989) stated that the influential data present important clues about the model and process under study. One approach that can detect influential data structures is to compare the estimates of the model parameters based on the sample with and without the data structures of interest using an algorithm (Ouwens et al., 2001). According to Chatterjee and Hadi (1986), influential factors can be classified into five groups which include measures based on residuals, prediction matrix, volume of confidence ellipsoids, influence functions, and partial influence. In this research, an algorithm has been developed to identify influencial factors that are based on influence functions. The identified influential factors can be used as guidelines or a basis for companies to strategize their plans for improving business performances.

In accessing a company's performance, many factors are considered. These include independent factors, dependent factors, and control factors (Maiga and Jacobs, 2003).

The contents of the thesis is for internal user only

References

Bergin, S. & Ronan, R. (2005). Programming: Factors that Influence Success. ACM1581139977/05/0002.

Byrd, T.A. & Turner, D.E. (2000). Measuring the Flexibility of Information Technology Infrastructure: Exploratory Study and Construct. *Journal of Management Information System*, Summer, pp.167-208.

Camison, C. & Villar-Lopez, A. (2010). Effect of SMEs' International Experience on Foreign Intensity and Economic Performance: The Mediating Role of Internationally Exploitable Assets and Competitive Strategy. *Journal of Small Business Management*, 48(2), p.116-151.

Chapman, P., Clinton, J., Kerber, R., Khabaza,T., Reinartz,T., Sherer, C. & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. pp78. Retrieved from the World Wide Web on November 02,2011, at http://www.markosweb.com/www/crisp-dm.org/.

Chatterjee, S. & Hadi, A.S. (1986). Influential Observation, High Leverage Points, and Outliers in Linear Regression. *Statistical Science*, Vol.1, No.3, pp.379-393.

Cook, R.D. (1979). Influential Observation in Linear Regression. *Journal of American Statistical Association*, Vol.74, No.365, pp.169-174.

Cox, J.L., Rothman, M.B. & Karron, D.B. (2003). Development and testing of an algorithm and its implementation to validate Digital Morse methohs for segmentation of image data. *Computer Aides Surgery*, Associate Member IEEE, pp.69-70.

Engelmen, C. (1965). *Mathlab: A Program for On-Line Machine Assistance in Symbolic Computations*. Proceedings Fall Joint Computer Coference.

Fayyad, U.M., Piatetsky, S. & Smyth, U. (1996). From Data Mining to Knowledge Discovering An Overview. *Advances in Knowledge Discovering and Data Mining*. AAAI Press/The MIT Press, Menlo Park, CA, pp.1-3.

Fernandez, Z. & Nieto, M.J. (2005). Internationalization Strategy of Small and Medium-Sized Family Businesses: Some Influential Factors. Family Firm Institute, Inc. VolXVIII, No.1.

Gray, J.B. (1989). On the use of regression diagnostics. *Journal of Royal Statistic Society, Series D (The Statistician)*, Vol.38, No.2, pp.97-105.

Guyon, Lemaire, Boulle, Dror, & Vogel. (2009). *Analysis of the KDD Cup 2009: Fast Scoring on a Large Orange Customer Database*, JMLR:Workshop and Conference Proceedings 7:1-22.

Houser, J. & Zong, L. (2007). The ARL Multi-Model Sensor: A research tool for target signature collection, algorithm validation, and emplacement studies. *US Army Research Laboratary*, IEEE.

Ignitia, M.J. & Irwan, B. (2004). Strategic Business-IT Alignment and Factors of Influence: A Case Study in a Public Tertiary Education Institution. *Proceeding of SAICSIT* pp.147-156.

Im, K.S., Dow, K.E. & Grover, V. (2001). A Reexamination of IT Investment and the Market Value of the firm: An Event Study Methodology. *Information System Research* Vol.12, No.I, pp.103-117.

Kevin, B.H. & Vinod, R.S. (2009). *An Empirical Analysis of the Effect of Supply Chain Disruption on Long-Run Stock Price Performance and Equity Risk of the Firm*. DOI: 10.1111/j.1937-5956.2005.tb00008.x.

Lawrence, K.D., Kudyba, S. & Klimberg, R.K. (2008). Data Mining Methods and Application. ISBN 0-8493-8522-9.

Maiga, A.S. & Jacobs, F.A. (2003). Organizatinal Effectiveness Analysis. ISSN: 1045-3695

Miller, T.W. (2005). Data and Text Mining: A Business Application Approach. ISBN 0-13-140085-1.

Morck, R. (2000). The information content of stock markets: why do emerging markets have synchronous stock price movements?. *Journal of Financial Economics*. Vol58, pp.215-260.

Nisbet, R., Elder, J. & Miner, G. (2009). Statistical Analysis and Data Mining. ISBN 978-0-12-374765-5.

Nuhanovic, A., Glavic, M. & Prljaca, N. (1998). Validation of a clustering algorithm for voltage stability analysis on the Bosnian electric power system. *IEE Proc-Gener.Transm.Distrib*, Vol.145, No 1.

Oleg, K., Yurij, S. & Oleksandra, M. (2011), *Comparison Analysis of Methods Implemented in MATHLAB for Fuzzy Logic Algorithms*, CAD/CAM Department, Lviv Polytechnic National University, 12, S.Bandery Str., Lviv, 79013, UKRAINE.

Oracle. (2005). Oracle @ Data Mining Concepts. *11g Release 1(11.1)*. Part number B28129-04

Osuna, R. (2002). Pattern Analysis. *CSCE 666, CSE@TAMU*.

Ouwens, M.J. (2001). Local Influential to Detect Influential Data Structures for Generalized Linear Mixed Models. *Biometrics, International Biometric Society*, Vol.57, No.4, pp.1166-1172.

Pattengale, N. (2010). Uncovering Hidden Phylogenetic Consensus in Large Datasets. IEEE/ACM Transaction on Computational Biology and Bioinformatics, Vol.X, No.X, X-X 201X 1

Pownall, G, Wasley, C. & Waymire, G. (1993). The stock price effects of alternative types of management earning forecasts. T*he Accounting Review*. Vol68, No.4.

Pratap, R. (1998). *Getting started with MATLAB 5 – A quick introduction for scientists and engineers.*. Oxford University Press. pp.240. ISBN-10: 0195129474. ISBN-13: 9780195129472

Ramadevi, Y. Rao, C.R. & Vivekchan, R. (2007). Decision tree Induction using Rough Set Theory-Comparative Study. *Journal of Theoretical and Applied Information Technology*, p.110-114.

Roger, G. (2000). *New Direction in Scientific Software Mathematics*. Scientific Computing and Instrumentation.

Shafer, J. Agrawal, R. & Metha, M. (1996), *SPRINT : A Scalable Parallel Classifier for Data Mining*, In Proceedings of the 22nd VLDB Conference, Bombay, India, pp.544-5555.

Suwardy, T., Ratnatunga, J., Sohal, A. & Speight, G. (2003). IT projects: evaluation, outcomes and impediments. *Benchmarking: An International Journal*. Vol.10:4, pp.325-342.

Tan, P.N., Steinbach, M. & Kumar, V. (2006). Introduction of Data Mining. ISBN 0-321-42052-7.

Li, T. & Ruan, D. (2007). An extended process model of knowledge discoring in database,. Vol 20, No.2.

Tlili, R. & Shamani, Y. (2011). *Executing Association Rule Mining Algorithms under a Grid Computing Envirinment*. ACM978-1-4503-0809-0/11/05.

Yang, W., Tan, B., Huang, D., Rautiainen, M., Shabanov, N.V., Wang, Y. Privette, J.L. (2006). MODIS Leaf Area Index Products: From validation to algorithm improvement. *IEEE Transactions on Geoscience and Remote Sensing*, Vol.44, No 7.